Viva La Correlación!

- Say X and Y are arbitrary random variables
  - Correlation of X and Y, denoted $\rho(X, Y)$:
    $$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var(X)Var(Y)}}}$$
  - Note: $-1 \le \rho(X, Y) \le 1$
  - Correlation measures <u>linearity</u> between X and Y
  - $\rho(X, Y) = 1 \quad \Rightarrow$ Y = aX + b where $a = \sigma_y/\sigma_x$
  - $\rho(X, Y) = -1 \quad \Rightarrow$ Y = aX + b where $a = -\sigma_y/\sigma_x$
  - $\rho(X, Y) = 0 \quad \Rightarrow$ absence of <u>linear</u> relationship
    - But, X and Y can still be related in some other way!
  - If $\rho(X, Y) = 0$, we say X and Y are "uncorrelated"
    - Note: Independence implies uncorrelated, but **<u>not</u>** vice versa!

Fun with Indicator Variables

- Let I_A and I_B be indicators for events A and B
  $$I_A = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases} \qquad I_B = \begin{cases} 1 & \text{if } B \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$
  - $E[I_A] = P(A)$, $E[I_B] = P(B)$, $E[I_A I_B] = P(AB)$
  - $\mathrm{Cov}(I_A, I_B) = E[I_A I_B] - E[I_A]\,E[I_B]$
    $= P(AB) - P(A)P(B)$
    $= P(A \mid B)P(B) - P(A)P(B)$
    $= P(B)[P(A \mid B) - P(A)]$
  - $\mathrm{Cov}(I_A, I_B)$ determined by $P(A \mid B) - P(A)$
  - $P(A \mid B) > P(A) \iff \rho(I_A, I_B) > 0$
  - $P(A \mid B) = P(A) \iff \rho(I_A, I_B) = 0$ (and $\mathrm{Cov}(I_A, I_B) = 0$)
  - $P(A \mid B) < P(A) \iff \rho(I_A, I_B) < 0$

Can't Get Enough of that Multinomial

- Multinomial distribution
  - n independent trials of experiment performed
  - Each trials results in one of m outcomes, with respective probabilities: $p_1, p_2, \ldots, p_m$ where $\sum_{i=1}^{m} p_i = 1$
  - X_i = number of trials with outcome i

$$P(X_1 = c_1, X_2 = c_2, \ldots, X_m = c_m) = \binom{n}{c_1, c_2, \ldots, c_m} p_1^{c_1} p_2^{c_2} \ldots p_m^{c_m}$$

  - E.g., Rolling 6-sided die multiple times and counting how many of each value {1, 2, 3, 4, 5, 6} we get
  - Would expect that X_i are negatively correlated
  - Let's see... when $i \ne j$, what is $\mathrm{Cov}(X_i, X_j)$?

Covariance and the Multinomial

- Computing $\mathrm{Cov}(X_i, X_j)$
  - Indicator $I_i(k)$ = 1 if trial k has outcome i, 0 otherwise
    $$E[I_i(k)] = p_i \qquad X_i = \sum_{k=1}^{n} I_i(k) \qquad X_j = \sum_{k=1}^{n} I_j(k)$$
  - $\mathrm{Cov}(X_i, X_j) = \sum_{a=1}^{n}\sum_{b=1}^{n} \mathrm{Cov}(I_i(b), I_j(a))$
  - When $a \ne b$, trial a and b independent: $\mathrm{Cov}(I_i(b), I_j(a)) = 0$
  - When $a = b$: $\mathrm{Cov}(I_i(b), I_j(a)) = E[I_i(a)I_j(a)] - E[I_i(a)]E[I_j(a)]$
  - Since trial a cannot have outcome i and j: $E[I_i(a)I_j(a)] = 0$

$$\mathrm{Cov}(X_i, X_j) = \sum_{a=b=1}^{n} \mathrm{Cov}(I_i(b), I_j(a)) = \sum_{a=1}^{n} (-E[I_i(a)]E[I_j(a)])$$
$$= \sum_{a=1}^{n} (-p_i p_j) = -np_i p_j$$ $\Rightarrow$ X_i and X_j negatively correlated

Multinomials All Around

- Multinomial distributions:
  - Count of strings hashed across buckets in hash table
  - Number of server requests across machines in cluster
  - Distribution of words/tokens in an email
  - Etc.
- When m (# outcomes) is large, p_i is small
  - For equally likely outcomes: $p_i = 1/m$
    $$\mathrm{Cov}(X_i, X_j) = -np_i p_j = -\frac{n}{m^2}$$
  - Large $m \Rightarrow X_i$ and X_j very mildly negatively correlated
  - Poisson paradigm still applicable

Conditional Expectation

- X and Y are jointly discrete random variables
  - Recall conditional PMF of X given Y = y:
    $$p_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$
- Define conditional expectation of X given Y = y:
  $$E[X \mid Y = y] = \sum_x x P(X = x \mid Y = y) = \sum_x x\, p_{X|Y}(x \mid y)$$
- Analogously, jointly continuous random variables:
  $$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \qquad E[X \mid Y = y] = \int_{-\infty}^{\infty} x\, f_{X|Y}(x \mid y)\, dx$$

Rolling Dice

- Roll two 6-sided dice D_1 and D_2
  - X = value of $D_1 + D_2$ Y = value of D_2
  - What is E[X | Y = 6]?

$$E[X \mid Y = 6] = \sum_x x P(X = x \mid Y = 6)$$

$$= \left(\frac{1}{6}\right)(7 + 8 + 9 + 10 + 11 + 12) = \frac{57}{6} = 9.5$$

  - Intuitively makes sense: 6 + E[value of D_1] = 6 + 3.5

Hyper for the Hypergeometric

- X and Y are independent random variables
  - X ~ Bin(n, p) Y ~ Bin(n, p)
  - What is E[X | X + Y = m], where m ≤ n?
  - Start by computing P(X = k | X + Y = m):

$$P(X = k \mid X + Y = m) = \frac{P(X = k, X + Y = m)}{P(X + Y = m)} = \frac{P(X = k, Y = m - k)}{P(X + Y = m)} = \frac{P(X = k)P(Y = m - k)}{P(X + Y = m)}$$

$$= \frac{\binom{n}{k} p^k (1-p)^{n-k} \cdot \binom{n}{m-k} p^{m-k} (1-p)^{n-(m-k)}}{\binom{2n}{m} p^m (1-p)^{2n-m}} = \frac{\binom{n}{k} \cdot \binom{n}{m-k}}{\binom{2n}{m}}$$

  - Hypergeometric: (X | X + Y = m) ~ HypG(*m*, *2n*, *n*)
  - E[X | X + Y = m] = *nm*/2*n* = *m*/2

total successes total trials “X” trials

Properties of Conditional Expectation

- X and Y are jointly distributed random variables

$$E[g(X) \mid Y = y] = \sum_x g(x) p_{X|Y}(x \mid y) \quad \text{or} \quad \int_{-\infty}^{\infty} g(x) f_{X|Y}(x \mid y)\, dx$$

- Expectation of conditional sum:

$$E\left[\sum_{i=1}^{n} X_i \mid Y = y\right] = \sum_{i=1}^{n} E[X_i \mid Y = y]$$

Expectations of Conditional Expectations

- Define *g*(Y) = E[X | Y]
  - *g*(Y) is a random variable
  - For any Y = y, *g*(Y) = E[X | Y = y]
    - This is just function of Y, since we sum over all values of X
  - What is E[E[X | Y]] = E[*g*(Y)]? (Consider discrete case)

$$E[E[X \mid Y]] = \sum_y E[X \mid Y = y] P(Y = y)$$

$$= \sum_y [\sum_x x P(X = x \mid Y = y)] P(Y = y)$$

$$= \sum_y \sum_x x P(X = x, Y = y) = \sum_x x \sum_y P(X = x, Y = y)$$

$$= \sum_x x P(X = x) = E[X] \quad \text{(Same for continuous)}$$

Analyzing Recursive Code

```
int Recurse()
{
   int x = randomInt(1, 3);  // Equally likely values

   if (x == 1) return 3;
   else if (x == 2) return (5 + Recurse());
   else return (7 + Recurse());
}
```

- Let Y = value returned by `Recurse()`. What is E[Y]?

$$E[Y] = E[Y \mid X = 1]P(X = 1) + E[Y \mid X = 2]P(X = 2) + E[Y \mid X = 3]P(X = 3)$$

$$E[Y \mid X = 1] = 3 \qquad E[Y \mid X = 2] = 5 + E[Y] \qquad E[Y \mid X = 3] = 7 + E[Y]$$

$$E[Y] = 3(1/3) + (5 + E[Y])(1/3) + (7 + E[Y])(1/3) = (1/3)(15 + 2E[Y])$$

$$E[Y] = 15$$

Random Number of Random Variables

- Say you have a web site: `PimentoLoaf.com`
  - X = Number of people/day visit your site. X ~ N(50, 25)
  - Y_i = Number of minutes spent by visitor *i*. Y_i ~ Poi(8)
  - X and all Y_i are independent
  - Time spent by all visitors/day: $W = \sum_{i=1}^{X} Y_i$. What is E[W]?

$$E[W] = E\left[\sum_{i=1}^{X} Y_i\right] = E\left[E\left[\sum_{i=1}^{X} Y_i \mid X\right]\right] = E[X \cdot E[Y_i]] = E[X]E[Y_i] = 50 \cdot 8$$

$$E\left[\sum_{i=1}^{X} Y_i \mid X = n\right] = \sum_{i=1}^{n} E[Y_i \mid X = n] = \sum_{i=1}^{n} E[Y_i] = nE[Y_i]$$

$$E\left[\sum_{i=1}^{X} Y_i \mid X\right] = X \cdot E[Y_i]$$

Conditional Variance

- Recall definition: Var(X) = $E[(X - E[X])^2]$
  - Define: $Var(X \mid Y) = E[(X - E[X \mid Y])^2 \mid Y]$
- Derived: Var(X) = $E[X^2] - (E[X])^2$
  - Can derive: $Var(X \mid Y) = E[X^2 \mid Y] - (E[X \mid Y])^2$
- After a bit more math (in the book):
  - $Var(X) = E[Var(X \mid Y)] + Var(E[X \mid Y])$
  - Intuitively, let Y = true temperature, X = thermostat value
  - Variance in thermostat readings depends on:
    - Average variance in thermostat at different temperatures +
    - Variance in average thermostat value at different temperatures

Making Predictions

- We observe random variable X
  - Want to make prediction about Y
  - E.g., X = stock price at 9am, Y = stock price at 10am
  - Let $g(X)$ be function we use to predict Y, i.e.: $\hat{Y} = g(X)$
  - Choose $g(X)$ to minimize $E[(Y - g(X))^2]$
  - Best predictor: $g(X) = E[Y \mid X]$
  - Intuitively: $E[(Y - c)^2]$ is minimized when c = E[Y]
    - Now, you observe X, and Y depends on X, then use c = E[Y | X]
  - You just got your first baby steps into Machine Learning
    - We'll go into this more rigorously in a few weeks

Speaking of Babies...

- Say my height is *X* inches (*x* = 71)
  - My son:  He does not look like: 
  - Say, historically, sons grow to heights Y ~ N(*X* + 1, 4), where *X* is height of father
    - Y = (X + 1) + C   where C ~ N(0, 4)
  - What should I predict for the eventual height of my son?
  - $E[Y \mid X = 71] = E[X + 1 + C \mid X = 71]$
    $= E[72 + C] = E[72] + E[C] = 72 + 0$
    $= 72$ inches